

Counting Eels and Iraqi Refugees

301108 Thinking About Data

WESTERN SYDNEY
UNIVERSITY



School of Computer, Data and Mathematical Sciences

Week 2 - Categorical Data



- 1 **Iraqi Refugees**
- 2 **Bar plots**
- 3 **Comparison to population**
- 4 **Eels - comparing two sets of counts**
- 5 **Chi square distributions**



- 1 **Iraqi Refugees**
- 2 **Bar plots**
- 3 **Comparison to population**
- 4 **Eels - comparing two sets of counts**
- 5 **Chi square distributions**



- In Uribe Guajardo et al. Int J Ment Health Syst (2016), the authors looked at the level of distress in 443 Iraqi refugees.
- The distress level is measured by a psychological instrument (known as the K10) and is classified as one of Low, Moderate, High or Very High distress.
- The following table was obtained

low	moderate	high	very high
123	70	93	157



- The Australian Institute of Health and Welfare (AIHW) also uses the K10 instrument to assess the Australian population.
- Using a very large sample (more than 10,000) they estimate the following;

low	moderate	high	very high
70.65	18.5	7.41	3.43

- Does the distribution of refugees distress differ from that of the Australian population?



- 1 Iraqi Refugees
- 2 **Bar plots**
- 3 Comparison to population
- 4 Eels - comparing two sets of counts
- 5 Chi square distributions

The first thing to do when confronted with data of this type is to draw a plot. Tables of numbers are harder to interpret.

- The type of plot that we would use for this data is a *bar plot*
- The data consists of counts in a series of categories
- The idea of a bar plot is to draw a bar or box, whose height represents the count, and draw one for each category



The first thing to do when confronted with data of this type is to draw a plot. Tables of numbers are harder to interpret.

- The type of plot that we would use for this data is a *bar plot*
- The data consists of counts in a series of categories
- The idea of a bar plot is to draw a bar or box, whose height represents the count, and draw one for each category

When to use barplots

Barplots are ideal for visualising the counts associated to a set of categories. In this case, we have the count of people associated to each distress level category.

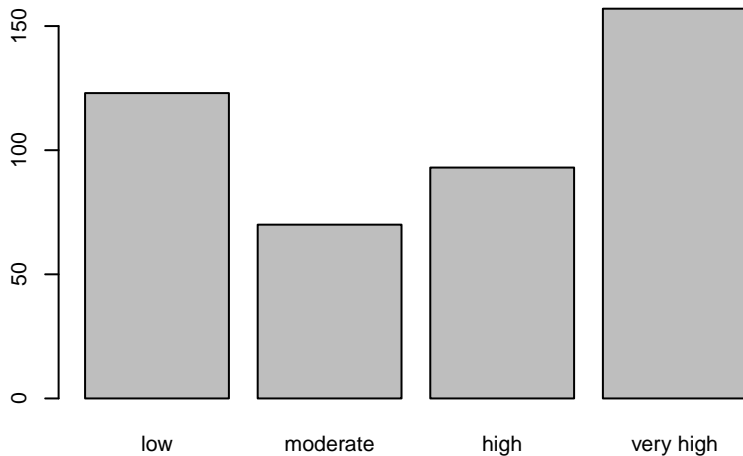


Figure: Barplot of numbers of refugees in each distress category

Bar plots can be drawn vertically or horizontally

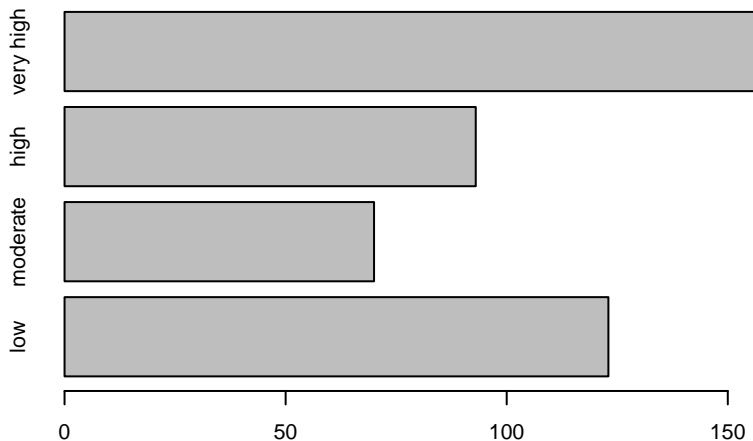


Figure: Horizontal barplot of numbers of refugees in each distress category

Bar plot for the AIHW data

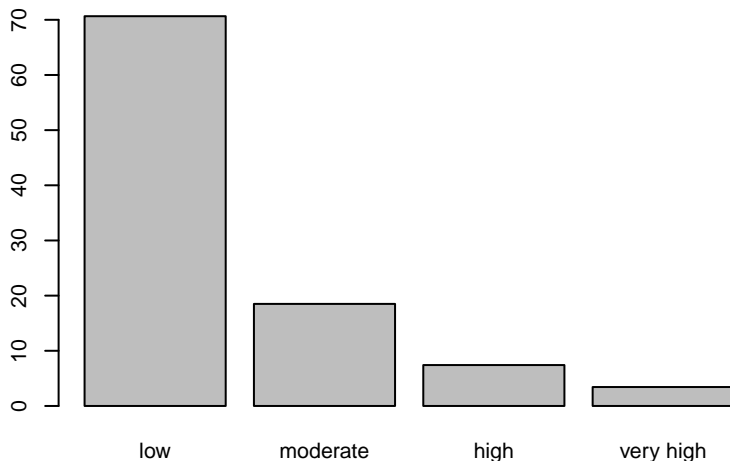


Figure: Barplot of the distress category proportions of the Australian population.

Bar plots can be combined to compare sets of data. The data from each data set should be either:

- counts of occurrences — an absolute comparison
- or percentages/proportions — when we are interested mainly in the distribution

For the refugee data we are mainly interested in whether a higher proportion of refugees are in certain distress categories than the Australian population. So we convert the counts to percentages (by dividing by the total and multiplying by 100)

Refugees v. the Australian population

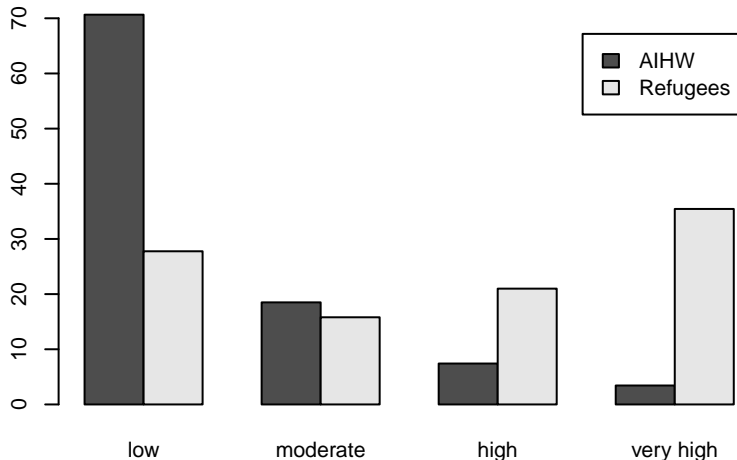


Figure: Comparing the distress categories of refugees and the Australian population.

Refugees v. the Australian population

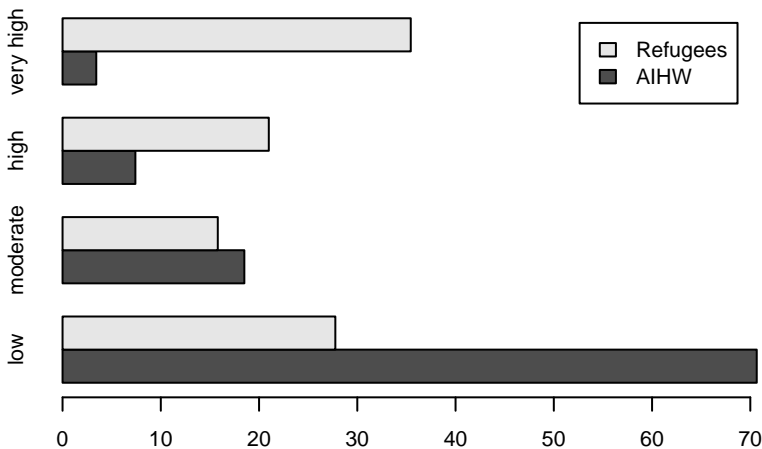


Figure: Comparing the distress categories of refugees and the Australian population.

- Sometimes stacked bars are used.

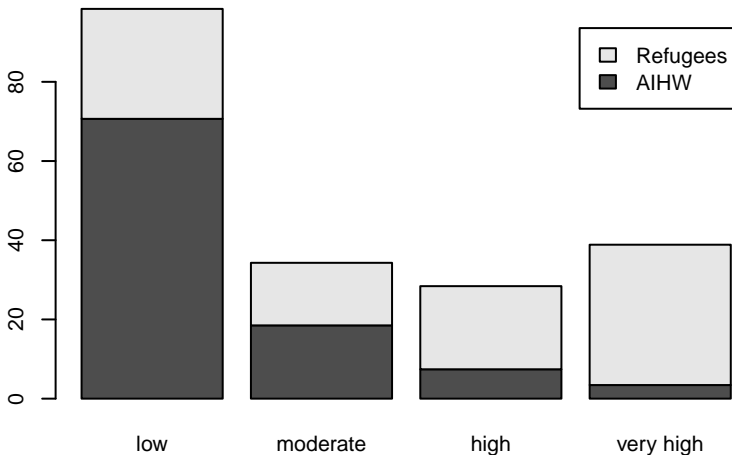


Figure: Comparing the distress categories of refugees and the Australian population.



Problem

Draw a barplot to examine if there difference in the eye colour distribution between genders.

	Brown	Blue	Hazel	Green
Male	98	101	47	33
Female	122	114	46	31



- 1 Iraqi Refugees
- 2 Bar plots
- 3 Comparison to population
- 4 Eels - comparing two sets of counts
- 5 Chi square distributions



Question of Interest

Does the distribution of a refugee's distress differ from that of the Australian population?

- Can we use a similar approach to that used for the digits of π ?
- Doing so means that we would need to randomly allocate the 443 individuals to the four distress levels, then compare it to our sample. This would allow us to check if all distress levels are equally likely (the distribution is uniform).



Question of Interest

Does the distribution of a refugee's distress differ from that of the Australian population?

- Can we use a similar approach to that used for the digits of π ?
- Doing so means that we would need to randomly allocate the 443 individuals to the four distress levels, then compare it to our sample. This would allow us to check if all distress levels are equally likely (the distribution is uniform).

But we are not interested in whether the distress level spread is uniform; we want to know if the distribution matches the AIHW percentages.

We ask “what would the refugee distress distribution look like if it matched the AIHW distribution?”

We need to simulate 443 individuals in the four categories so that the chances of being in each category are **not** uniform, but according to the AIHW percentages.

low	moderate	high	very high
70.65	18.5	7.41	3.43

For this example, we will round the numbers to simplify the simulation (if using a computer, this simplification is not needed).

low	moderate	high	very high
71	19	7	3

If we generate a random number (x say) between 1 and 100, and call the generated category:

- low if x is less than 71
- medium if x is between 72 and 90 ($=71+19$) (inclusive)
- high if x is between 91 and 97 ($=90+7$)
- very high if x is greater than 97

Repeating this, then on average, 71 out of 100 will be low, 19 out of 100 will be medium etc.

This approach can be extended to non-integer percentages. R does this for us, (Excel does not as far as I can tell).



Once we simulate the category counts, assuming that the AIHW percentages are true, we must ask “Are the category counts from the sample (shown below) consistent with the simulated counts from the AIHW proportions?”

low	moderate	high	very high
123	70	93	157



Here are the simulation results repeated ten times.

low	moderate	high	very high
292	99	36	16
313	86	31	13
311	92	23	17
292	100	29	22
320	80	31	12
311	91	30	11
310	85	34	14
328	74	30	11
313	80	33	17
304	93	33	13

In this case, we have 443 individuals in four categories, but they are not evenly spread.

- The AIHW has 70.65% in the low category
- So the expected count in this category is $443 \times 70.65/100 = 312.99$, since the sample size is 443.

Below are the remaining expected counts when using a sample size of 443:

	low	moderate	high	very high
percent	70.65	18.50	7.41	3.43
expected	312.99	81.96	32.84	15.20



We need to measure how different the category counts are from their expected value.

- So, as before, we can subtract the expected value from the actual counts and square and sum.
- We can do this for the simulated data from the AIHW proportions *and* the actual Refugee counts.

Refugees = 59966.79

741.69	24.56	204.92	827.19	66.57	111.42	20.98	314.28	7.11	207.63
--------	-------	--------	--------	-------	--------	-------	--------	------	--------

A better distance

- It turns out that this distance does not give enough **weight** to differences where there a small expected counts versus where there are large expected counts.
- A better distance is to take the count minus the expected count squared **divided** by the expected **then** add up.

In maths notation,

$$\sum_i \frac{(O_i - E_i)^2}{E_i}$$

This is called the *chi-square distance* (or χ^2)



If you haven't seen things like

$$\sum_i \frac{(O_i - E_i)^2}{E_i}$$

before.

- O_i stand for the observed counts for each i , eg O_1 is the first observed count.
- E_i stands for the corresponding expected count
- \sum_i means sum or add up over i



	low	moderate	high	very high
refugees	123.00	70.00	93.00	157.0
expected	312.99	81.96	32.84	15.2

$$\frac{(123 - 312.99)^2}{312.99} + \frac{(70 - 81.96)^2}{81.96} + \frac{(93 - 32.84)^2}{32.84} + \frac{(157 - 15.2)^2}{15.2}$$
$$= 1550.08$$

Chi-square for Simulated Counts

Problem

Compute the χ^2 distance between one of the simulated sets of category counts and the expected set of category counts.

	low	moderate	high	very high
simulated	292.00	99.00	36.00	16.0
expected	312.99	81.96	32.84	15.2



Finally, we examine the χ^2 distance for all simulated counts, and examine if the χ^2 distance for the data looks like these.

Number of sets	Maximum chi-squared difference
10	8.54
100	10.79
1000	16.72
10000	21.68

The distance for refugees is much larger than any of these. This indicates that Iraqi refugee distress levels are probably not the same as the Australian population.



- 1 Iraqi Refugees
- 2 Bar plots
- 3 Comparison to population
- 4 Eels - comparing two sets of counts
- 5 Chi square distributions

Two species of eels are observed at three different habitats. The following counts are made.

	Border	Grass	Sand
G.moringa	264	127	99
G.vicinus	161	116	67

Are the distribution of Eel species the same?

Two species of eels are observed at three different habitats. The following counts are made.

	Border	Grass	Sand
G.moringa	264	127	99
G.vicinus	161	116	67

Are the distribution of Eel species the same?

This problem is different to the Iraqi Refugees data, since we do not have a distribution to compare to (for the refugee data, we used the distribution from the AIHW).

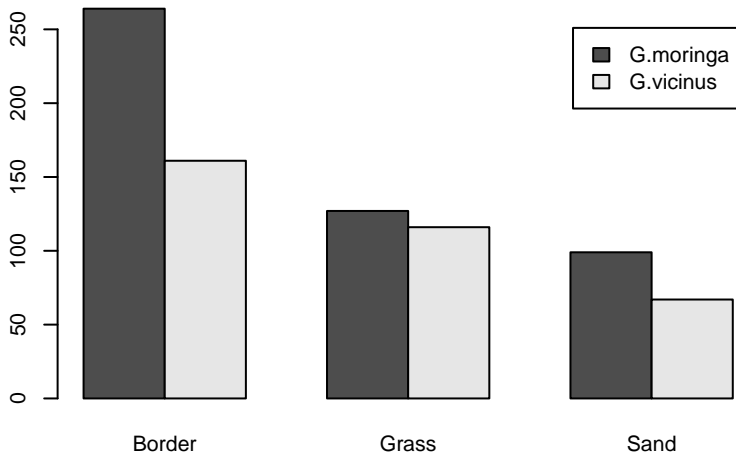


Figure: Eel counts at given habitat.

There certainly seems to be:

- a. Fewer *G.vicinus* overall
- b. A lot fewer *G.vinicus* in the Border habitat

Do we care about the overall number? Or just the spread/distribution?

Proportion of sightings in each habitat

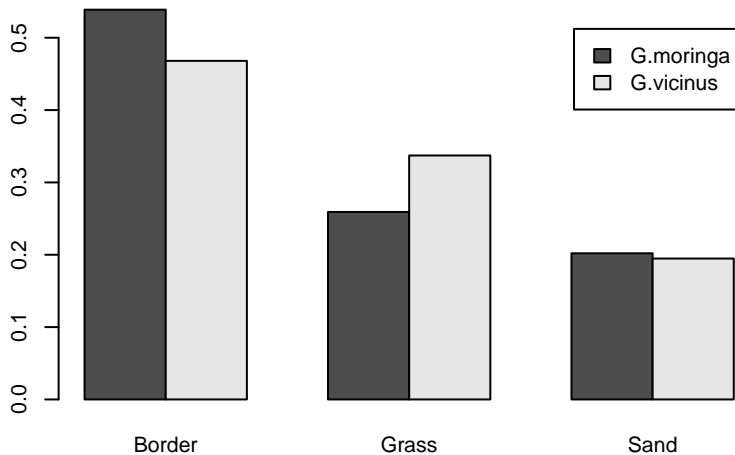


Figure: Eel proportions at given habitat.

This looks a bit like the refugees question, but it is subtly different.

- Previously, we asked if the distress level of refugees were different from the Australian population. And we essentially knew the average distress levels for the Australian population from a very large sample
- In this case, neither set of counts is all that large.
- To work as before we have to find an expected value for each count and a way to simulate counts, under the assumption that the species distributions are the same.



If the species distributions are the same, then we can aggregate the counts across species to estimate an overall habitat distribution

	Border	Grass	Sand
counts	425.00	243.00	166.0
percent	50.96	29.14	19.9

- So if the distribution is the same we **expect** to see the same percentage of each species count in each habitat.
- If there are $n_1 = 490$ eels of species G.moringa, then we should see $n_1 \times p_1$ in the Border habitat where p_1 is the proportion in Border,
- and $n_1 \times p_2$ in Grass where p_2 is the proportion in Grass etc.

The same applies to *G.vinicus*, with n_1 replaced with $n_2 = 344$

The data is:

	Border	Grass	Sand	Total
<i>G.moringa</i>	264	127	99	490
<i>G.vinicus</i>	161	116	67	344

The expected counts are:

	Border	Grass	Sand
<i>G.moringa</i>	249.7	142.77	97.53
<i>G.vinicus</i>	175.3	100.23	68.47

We need to simulate what the counts would look like if the two eel species shared the same distribution across habitats. To do this, for each of the 834 eels:

- we sample a species using the proportions from the data:

G.moringa	G.vicinus
0.58753	0.41247

- and sample a habitat using the proportions from the data:

Border	Grass	Sand
0.5095923	0.2913669	0.1990408



- We would then proceed by calculating the chi-squared distance from expected for our actual data and for a (large) number of simulations and compare.
- Previously we were comparing to an (essentially) fixed set of expected values, but now we used the proportions in the data to compute them.
- So we must recompute the expected values for every simulation



Problem

One simulation provided the counts below. Compute the χ^2 distance to the expected counts.

	Border	Grass	Sand
G.moringa	267	143	101
G.vicinus	173	86	64



Number of sets	Maximum chi-squared difference
10	6.08
100	10.71
1000	14.23
10000	17.62

The chi-square for the **actual** eel counts is 6.26

Its bigger than the max from 10 but not for a 100... Hmmm.

p-values

Lets look at this in more detail.

If we take 1000 of the simulated numbers (Im not going to print then here), it turns out that only 46 are greater than 6.26.

So although the eels are not further from expected than 1000 random simulations, they are than most of them.

Only 46 out of 1000 a further. Or 4.6%

This is called a *p*-value.

The idea here that “the distribution of eels across habitats does not differ” is called the *null hypothesis*.

The converse, that these distribution do differ, is the *alternative hypothesis*.

p value

The *p*-value is the **chance** or proportion of the time that we would see a chi-squared distance as large or larger than the actual distance for the data, if we simulate assuming the null hypothesis is true



There is an issue here, in that we have assumed a fixed number of eels were observed, but that the number in each species and in each habitat was not fixed.

Sometimes in two-way table like this might have some more fixed numbers.

Example 1: One fixed margin

- Suppose we are interested in the voting habits of Men versus Women.
- We find 500 men and 500 women and ask would they vote LNP, ALP or Green.
- In this design, the number of men and women is fixed in advance.

It is said to have one fixed **margin**

	ALP	Green	LNP
Men	177	120	203
Women	196	104	200

Example 2: Two fixed margins

- A wine taster is presented with 20 glasses of wine
- Ten are cool climate and ten warm climate wines.
- The taster is told there are ten of each and asked to say, by taste alone, which are the cool climate ones.

This table has two fixed margins - both rows and columns add up to 10

	Cool	Warm
Cool	8	2
Warm	2	8

When using the simulation approach, the simulation should follow the sampling strategy



- 1 **Iraqi Refugees**
- 2 **Bar plots**
- 3 **Comparison to population**
- 4 **Eels - comparing two sets of counts**
- 5 **Chi square distributions**

Chi square distributions

So far, the hypothesis testing process has been:

- 1 Compute a summary statistic (mean, median, χ^2 distance) of the sample data.
- 2 Generate many simulations of the data, where the Null Hypothesis is true.
- 3 Compute the summary statistic of each simulated data to obtain a distribution of summary statistics where the Null Hypothesis is true.
- 4 Compare the data statistic to the simulated data statistic. If they look different, then it is likely that the Null Hypothesis is false.

When using the χ^2 distance, we can approximate steps 2 and 3 using the χ^2 distribution.

Chi square distributions

- The Chi-squared distribution (sometimes written χ^2 -distribution) allows us to say approximately how typical a chi-squared distance is without simulating
- It lets us compute the proportion of times a value is exceeded if the **null** hypothesis is true
- It is actually a family of distributions, that depend on a parameter related to the number of counts being considered
- It requires the extra assumption that **the expected cell counts are not too small** (at least 5).

Chi square distributions

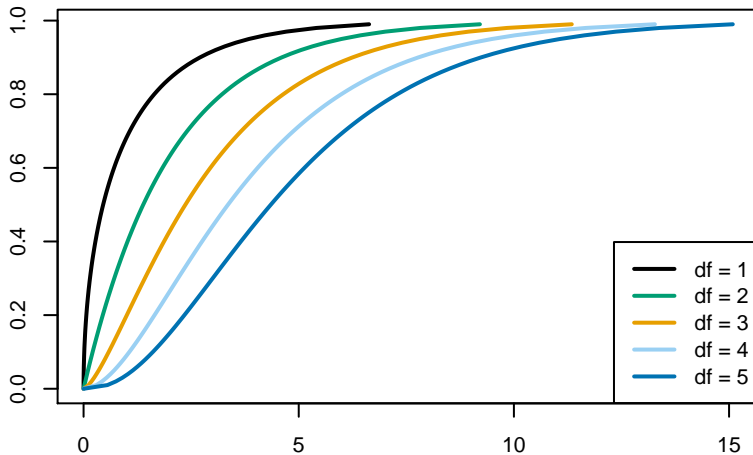


Figure: Chi square distributions with different degrees of freedom (df).

Chi square test

- For the first example — comparing to a known distribution — the chi-square distribution df to use is *One less than the number of cells in the count table* — 3 for iraqi refugees
- The actual chi-square distance was 1550.08. Using the graph (or tables or a command in R/excel), the proportion less than this would be virtually 1. So the proportion greater (the p-value) would be virtually zero.

Chi square test

- For the second example — comparing two sets of counts — the chi-square distribution df to use is *One less than the number of rows in the count table multiplied by one less than the number of columns in the count table* $(r - 1) \times (c - 1) = 2$ for the eels example
- The actual chi-square distance was 6.26. Using the graph (or tables or a command in R/excel), the proportion less than this would be over 0.9 (90%). So the proportion greater would be under 10%.
- In fact, the proportion greater (p-value) is 0.0437
- (We got $46/1000 = 0.046$ by simulation)



- Categorical data can be visualised using boxplots or tables.
- To compare if two samples of categorical data come from the same population or follow a given population, we can use simulation or a Chi square distribution.
- The simulation and degrees of freedom depends on how the experiment was run.